微信之道－至简

广州研发部　harveyzhou

腾讯大讲堂 http://djt.qq.com

VISIT…

关于我

- 周颢（harveyzhou）
- 2001年毕业于华南理工大学，计算机专业硕士
- 2005年加入腾讯广州研发部
- 历任QQ邮箱架构师，广研技术总监，T4技术专家，微信中心助理总经理

关于微信

- 移动互联网的探索者
- 10个月5000万手机用户
- 创造移动互联网用户增速的记录
- 千万级在线
- 苹果中国区AppStore月下载量第一
- 摇一摇每天次数过亿
- 腾讯战略级产品

微信的历程

腾讯大讲堂 http://djt.qq.com

微信的三位一体

- 产品的精准
- 项目的敏捷
- 技术的支撑

产品的精准—用简单规则构造复杂世界

- 张小龙，腾讯副总裁，广研灵魂人物
- 从第二代程序员旗手，到领军者，到产品传奇人物
- 从**Foxmail**，到**QQ**邮箱，到微信

微信的三位一体

- 产品的精准
- 项目的敏捷
- 技术的支撑

什么是敏捷

- 项目管理的技巧？**Scrum**？
- 矿工？

项目的敏捷

- 敏捷就是试错法
- 敏捷是一种态度
  - 产品决策是成功的第一因素
  - 允许发布前十分钟的变更
  - 给予产品决策以最大自由度

敏捷的困境

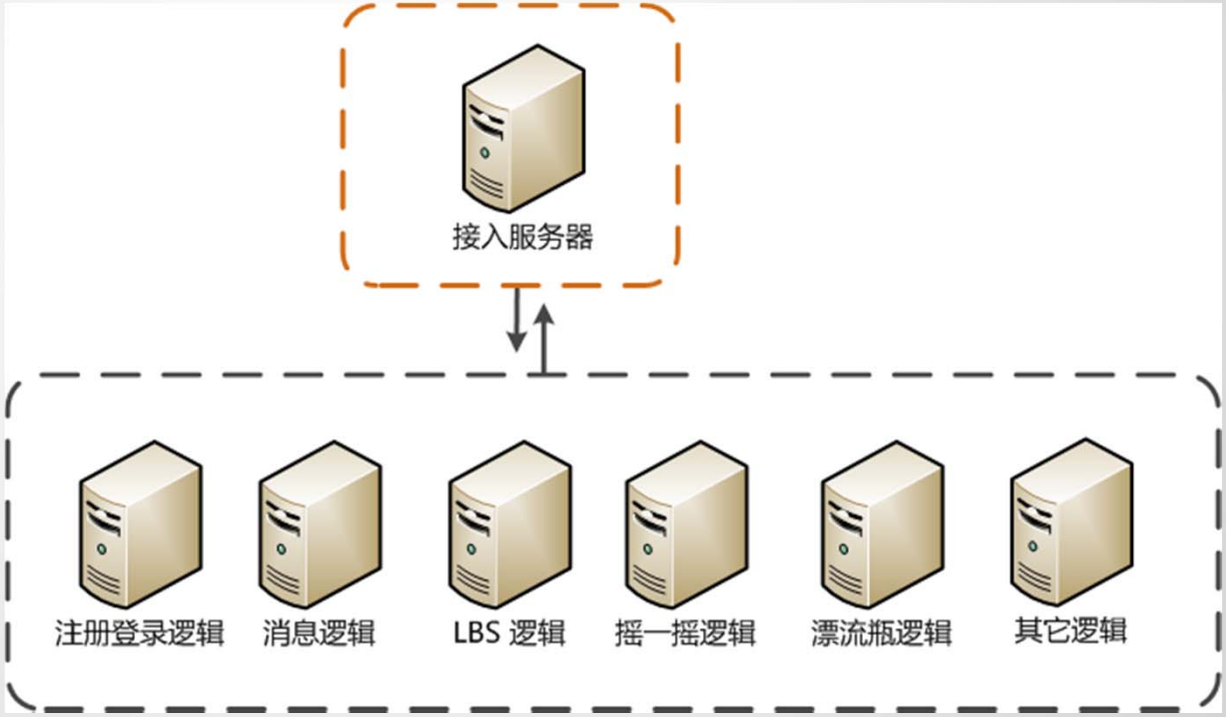
- 海量系统的复杂度
  - 千万级同时在线
  - 亿级摇一摇
  - 单集群百亿级服务请求
  - **99.95%**的可用性
- 海量系统上的敏捷，无异于悬崖边的跳舞

让敏捷变得简单

- 狂热的信念，**You can do it !**
- 稳固的技术支撑
  - 大系统小做
  - 让一切可扩展
  - 要有基础组件
  - 轻松的上线
    - 灰度，灰度，再灰度
    - 精细的监控
    - 迅速的响应

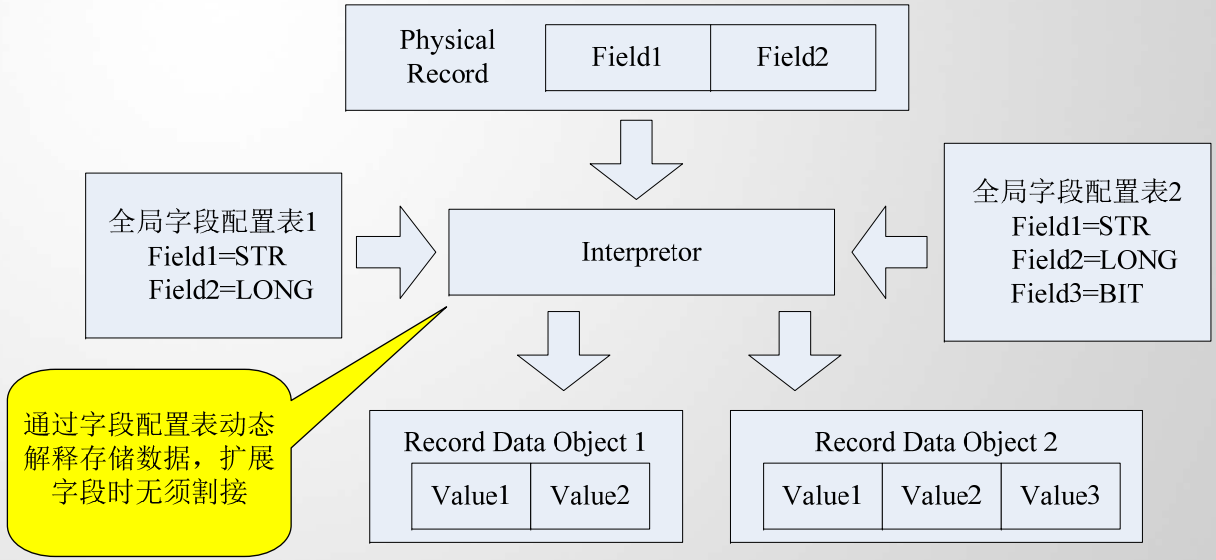
大系统小做

- 从代码分模块，到分离部署
- 灵活的折衷：混搭模式（重要/复杂逻辑分离，其它混合部署）



让一切可扩展

- 网络协议可扩展
  - **XML描述**
  - 向前兼容
  - 代码自动生成（**ProtocolBuffer & TLV**）
- 数据存储可扩展
  - **KV or TLV**
  - 字段配置表
  - 类SQL处理



要有基础组件

- **Svrkit：Client/Server自动代码生成框架**
  - **10分钟搭建内部服务器**
- **LogicServer：逻辑容器**
  - 随时添加新逻辑
- **OssAgent：监控/统计框架**
  - 所见即所得的监控报表
- 存储组件
  - 屏蔽容灾/扩容等复杂问题

灰度，灰度，再灰度

模块：mmsynclogicsvr
若找不到模块或服务器，请先找运维负责人，确认服务器状态是运营中
搜索：
已选5台服务器
还可选●台服务器
首页 上页 1 2 3 4 5 下页 末页
mmsynclogicsvr1
mmsynclogicsvr2
mmsynclogicsvr3
mmsynclogicsvr4
mmsynclogicsvr5
mmsynclogicsvr6
mmsynclogicsvr7
mmsynclogicsvr8
mmsynclogicsvr9
mmsynclogicsvr10
mmsynclogicsvr11
mmsynclogicsvr12
mmsynclogicsvr13
mmsynclogicsvr14
mmsynclogicsvr15
mmsynclogicsvr16
mmsynclogicsvr17
mmsynclogicsvr18
mmsynclogicsvr19
mmsynclogicsvr20
mmsynclogicsvr21
mmsynclogicsvr22
mmsynclogicsvr23
mmsynclogicsvr24
mmsynclogicsvr25
>>
<<
mmsynclogicsvr1
mmsynclogicsvr2
mmsynclogicsvr3
mmsynclogicsvr4
mmsynclogicsvr5
总计：●台，已灰度：21台
后台单至少灰度两次，且50%以上主机灰度，才能全面上线
全选所有
一次选够
灰度上线
返回
清空所有

You can do it !

>20个后台变更/天

每页显示 20 条　　首页 上页 1 2 3 4

单号	模块名	建单人	测试人	任务名	结束时间	状态
qqmail-120326-1654-0920	mmsynclogicsvr	bierhuang		mmsynclogics	2012-03-26 17:15:54	灰度完成
qqmail-120326-1635-4059	mmhdiconsvr	eddyzeng		mmhdiconsvr	2012-03-26 16:55:00	全上完成
qqmail-120326-1628-3189	ALL	paulohuang		mmlangconf	2012-03-26 16:39:18	全上中断[完成率:99.98%]
qqmail-120326-1627-3418	mmsynclogicsvr	bierhuang		mmsynclogics	2012-03-26 16:46:18	灰度回退完成
qqmail-120326-1605-5730	mmiconsvr	eddyzeng		mmiconsvr 修改	2012-03-26 16:45:42	全上完成
qqmail-120326-1544-2510	ALL	francowu		mmasyncmq7切走	2012-03-26 15:53:36	全上中断[完成率:99.98%]
qqmail-120326-1542-2150	mmopenappinfo	berryxie		mmopenappinf	2012-03-26 15:44:00	全上完成
qqmail-120326-1538-2239	ALL	junechen		mmadpush_cli	2012-03-26 15:54:59	全上中断[完成率:99.96%]
qqmail-120326-1536-2079	mmimglogicsvr	lynncui		mmimglogicsv	2012-03-26 15:38:21	全上完成
qqmail-120326-1532-2741	mmbottlelogicsvr	andrewang		更新 bottlelog	2012-03-26 16:53:04	全上完成
qqmail-120326-1528-2676	mmproxy	sharonzhang		mmproxy_cgi.	2012-03-26 15:34:43	灰度完成
qqmail-120326-1507-2070	mmbizindex	sharonzhang		mmbizindex上线	2012-03-26 15:10:33	全上完成
qqmail-120326-1505-4735	mmlogcenter	mariohuang		mmlogcenter重	2012-03-26 15:10:32	灰度中断[完成率:0.00%]
qqmail-120326-1412-1569	mmnewaddrbook	delphiliu		kvsvr 变更	2012-03-26 14:13:08	已建单
qqmail-120326-1411-3727	mmuserattr	stevenshe		kvstoreproxy	2012-03-26 14:30:27	全上完成
qqmail-120326-1410-0980	mmqqmsg	tommytang		配置QQ尾巴一条一条	2012-03-26 14:20:53	全上完成
qqmail-120326-1354-0514	mmadpush	junechen		mmadpush	2012-03-26 13:54:54	全上完成
qqmail-120326-1347-5984	mmlbslogicsvr	junechen		mmlbslogicsv	2012-03-26 17:13:40	全上完成

腾讯大讲堂 http://djt.qq.com

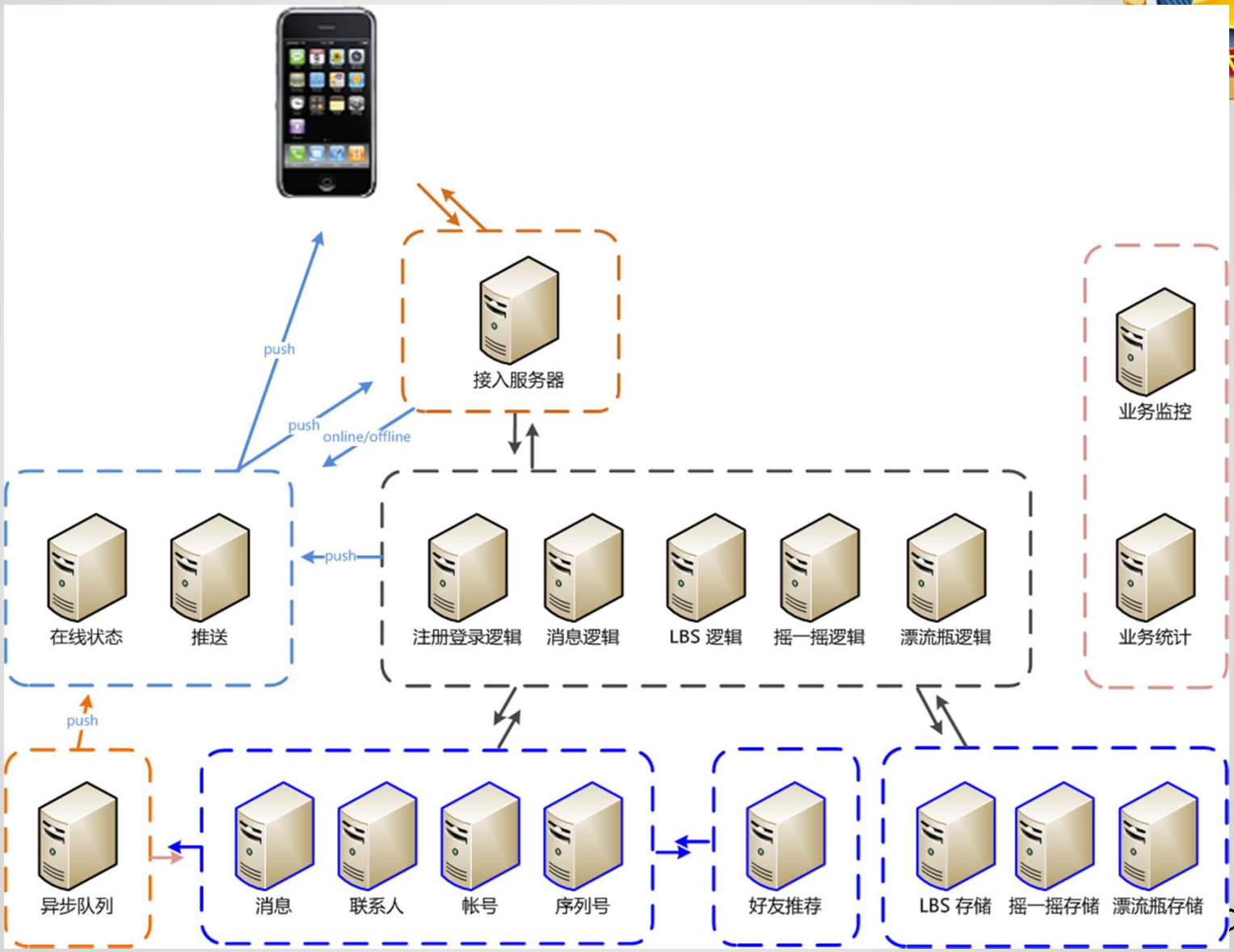
微信的三位一体

- 产品的精准
- 项目的敏捷
- 技术的支撑

技术的支撑—剥离复杂，让剩下的更简单

- 孙子兵法：古之所谓善战者，胜于易胜者也

微信架构


push
push
online/offline
接入服务器
业务监控
push
在线状态
推送
注册登录逻辑
消息逻辑
LBS 逻辑
摇一摇逻辑
漂流瓶逻辑
业务统计
push
异步队列
消息
联系人
帐号
序列号
好友推荐
LBS 存储
摇一摇存储
漂流瓶存储

q.com

关注复杂点

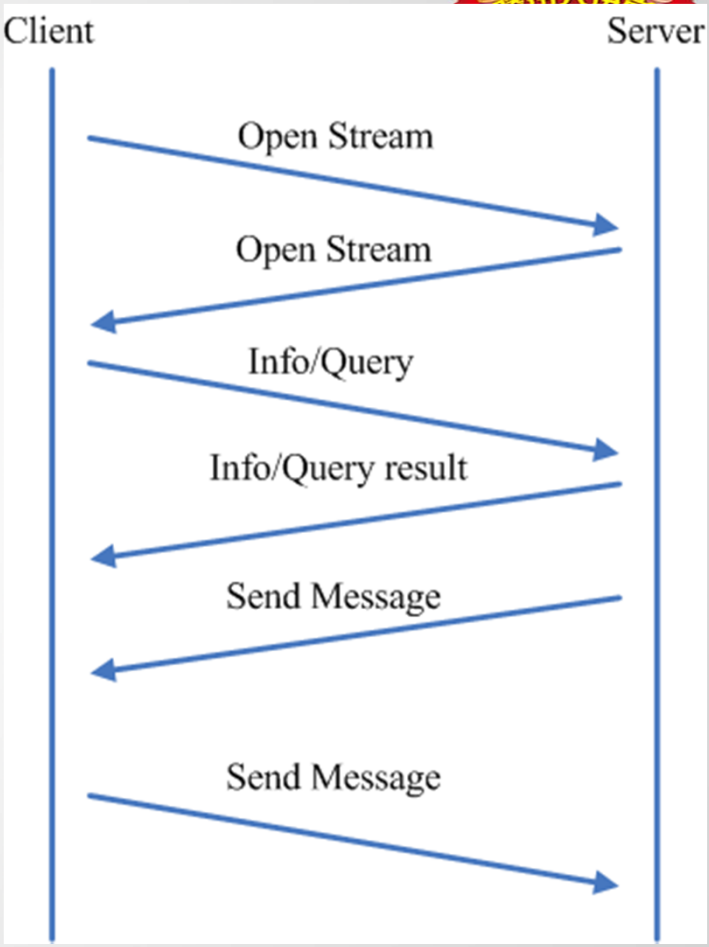
- 协议
- 容灾
- 轻重
- 监控

移动互联网的复杂性

- **CMWAP vs. CMNET**
- 在线 **vs.** 离线
- 连接不稳定
- 资费敏感
- 高延迟

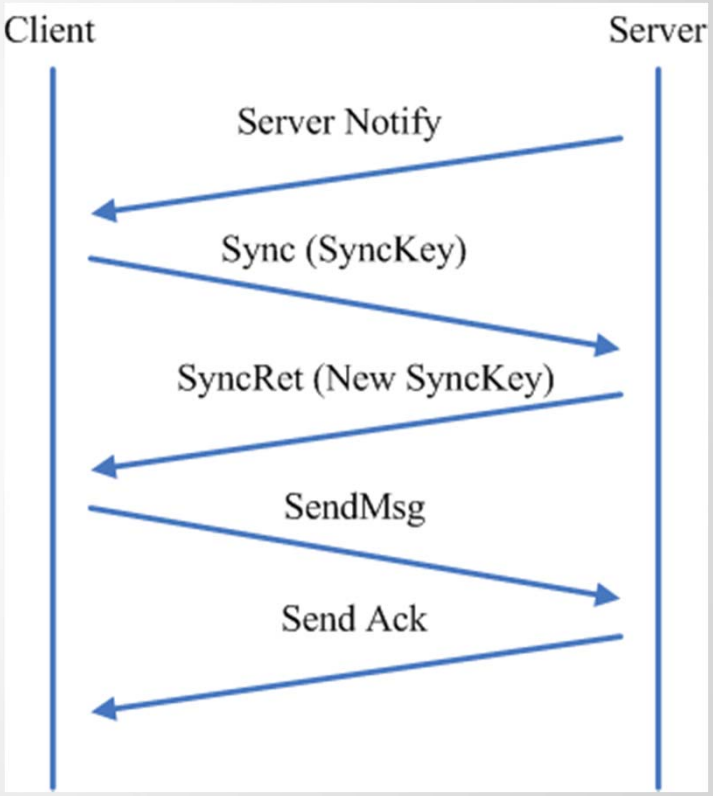
业界标准方案

- **Messaging And Presence Protocol**
  - **XMPP**
  - **SIP/SIMPLE**
- 优点：
  - 简单，大量开源实现
- 缺点
  - 流量大：状态初始化
  - 消息不可靠
- 把简单留给自己，把复杂留给别人？



SYNC协议

- 参考**ActiveSync**
- 状态同步：**Sync by SyncKey**
- 模式简化：**Notify & Client Pull**
- 实现更复杂，但 …



让剩下的更简单

- 简化交互模式
- 最小增量传输
- 最优重传控制
- **More important：** 消息可靠传输 **&** 按序到达

- *和菜头 @ 2011/11/26*
  比它炫的没它简单，
  比它简单的没它快，
  没有谁比它更快，
  哪怕在GPRS下，
  微信也能把进度条轻易推到底。

关注复杂点

- 协议
- 容灾
- 轻重
- 监控

在容灾之前－面向最坏的思考

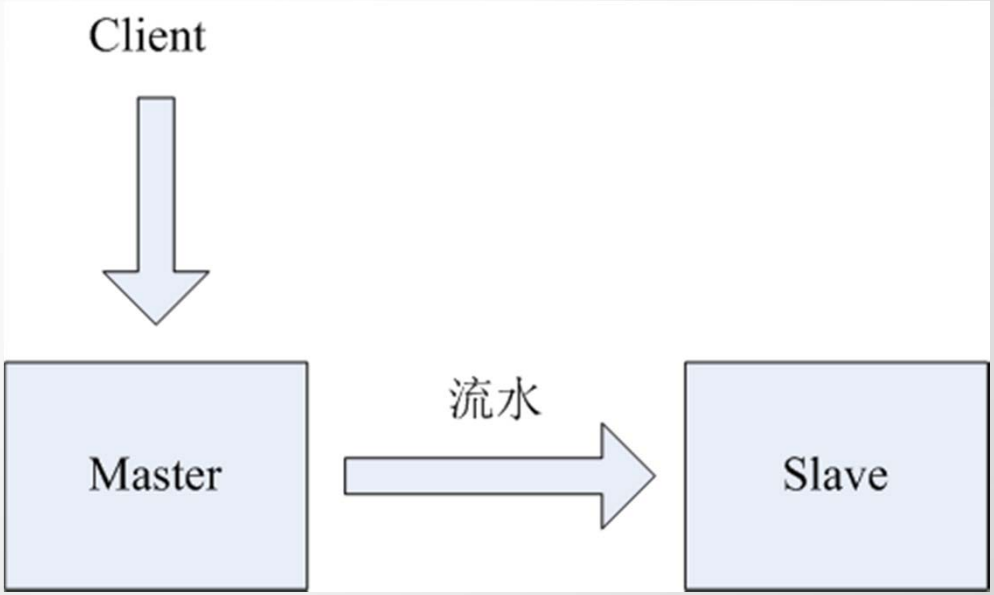
- 如果真的挂了 …. …
  - 防止雪崩，避免蝴蝶效应
    - 把防雪崩内置到组件
  - 柔性可用，追求不完美
    - 只求完美的团队，
      不能胜任海量服务。
      0/1完美 = **60**分
  - 保护点前置，赢得处理空间
    - 终端配合的容灾

存储层容灾－分而治之

- 与接入层/逻辑层对比
  - 接入层：**GSLB, LVS, IP redirect, Client Retry**
  - 逻辑层：无状态设计
- 存储层容灾是海量系统最复杂的设计
- 分而治之
  - 分离业务场景，寻求简单设计

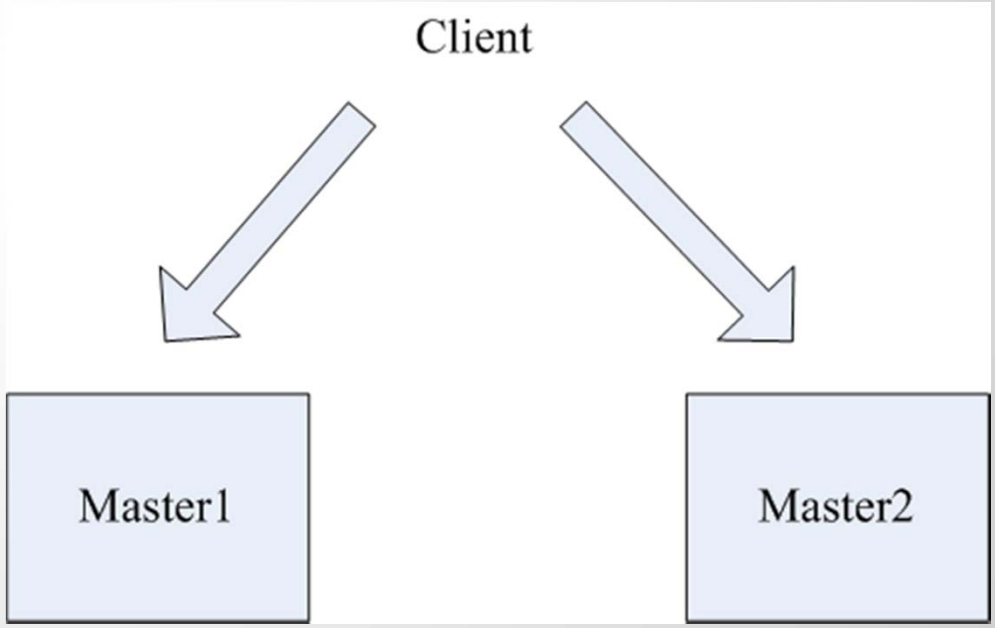
主备

- 实现简单
- 局限
  - 容忍最终一致性
  - 故障时不可写
- **Example**
  - 帐号系统



双写

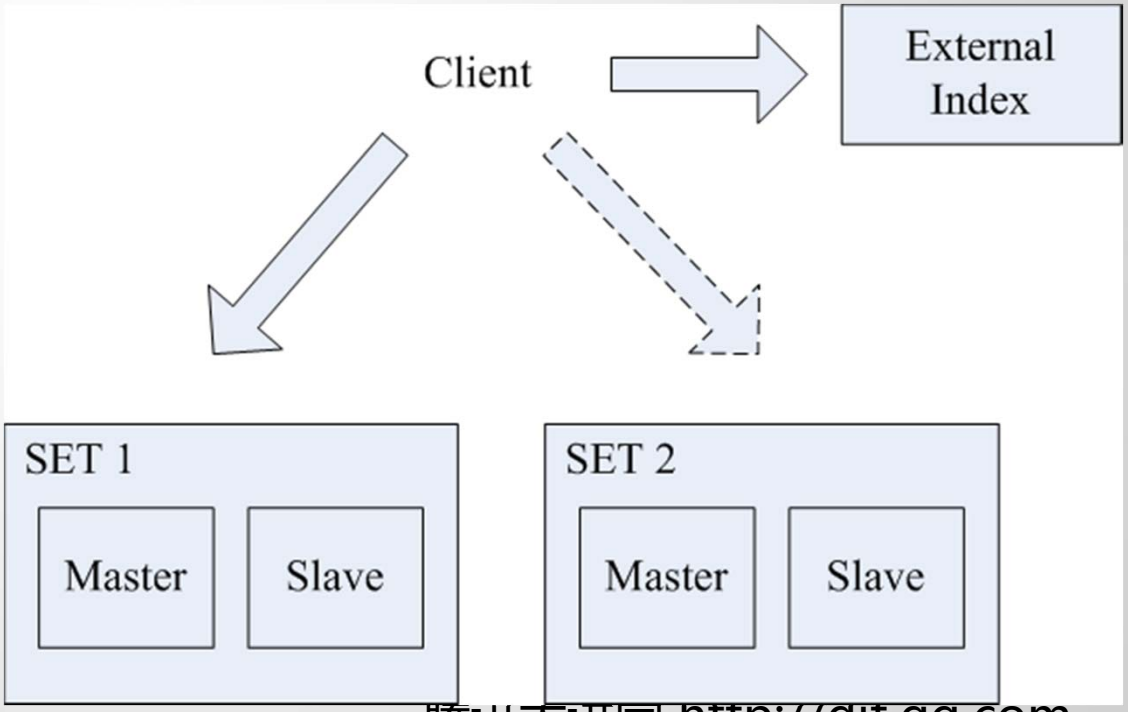
- 实现简单
- 故障时可写
- 局限
  - 容忍轻度数据丢失
- **Example**
  - 用户终端类型记录



SET模型＋双写

- 实现简单（注：**SET1**写入不成功时，切换到**SET2**写入）
- 完全一致的备份副本
- 局限
  - 只支持追加写
  - 需要外部索引
- 简化版**Google FS**

- **Example**
  - 语音/图片存储



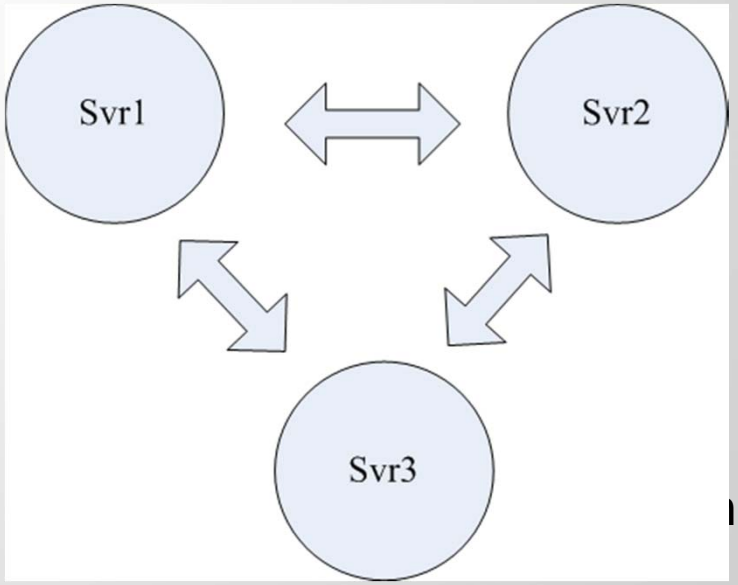
Quorum

- 分布式理论
  - **CAP理论**
  - **Paxos**
    - **Leslie Lamport**
    - **Chubby，ZooKeeper**
  - **Quorum：Amazon Dynamo**
    - **R+W>N**
    - **Vector Clock：解决冲突**
    - **Merkle Tree：节点恢复**

Simple Quorum

- 实现SYNC协议的序列发生器
- 极高稳定度要求

- 避免Vector Clock：递增的序列号
- 避免Merkle Tree：全量加载
  - 一位毕业生的创意：按SET分布，全量数据从2G减到200K



关注复杂点

- 协议
- 容灾
- 轻重
- 监控

终端的迷思

- 复杂的逻辑
- 高昂的变更
- 致命的风险

前轻后重

- 功能点后移，发挥后台快速变更的优势

接入优化：从GSLB到IP重定向

“偷流量”防御：屏蔽流量异常的终端

后台适配

- 群聊的例子
- 二维码扫描的例子

腾讯大讲堂 http://djt.qq.com

关注复杂点

- 协议
- 容灾
- 轻重
- 监控

监控的痛苦

- 海量的日志：数百G/小时
- 实时的图表：1分钟
- 灵活的需求：复杂的关联统计

- 鱼与熊掌不可兼得也

分而治之

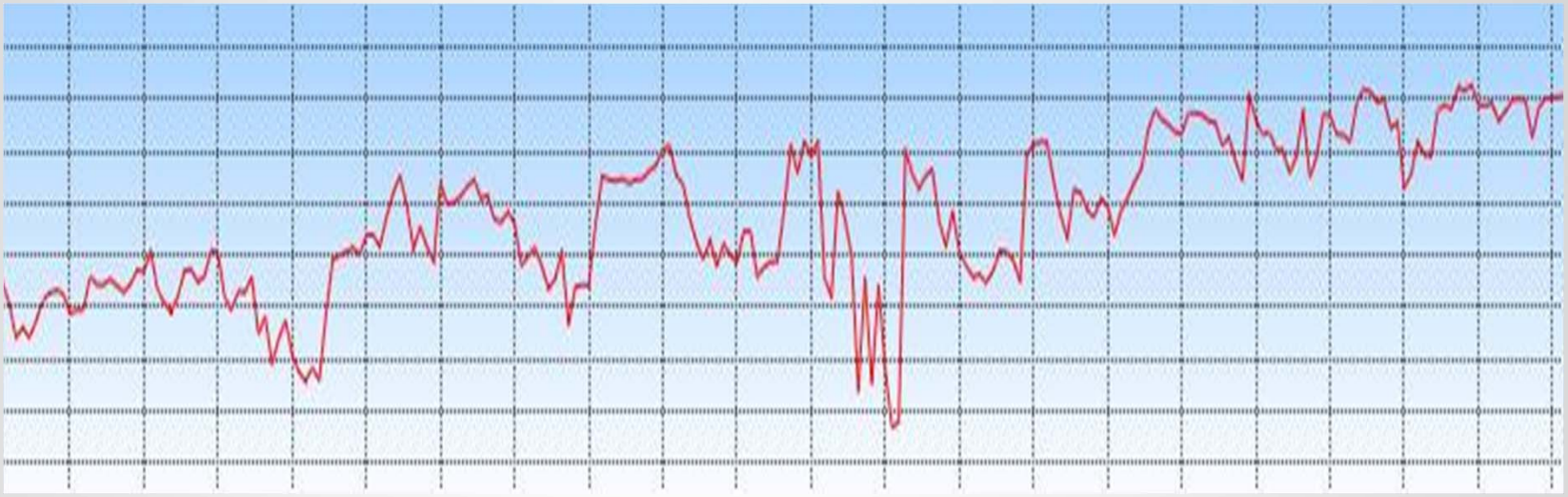
- 监控 != 统计
- 监控
  - 反映系统运行状态
  - 关联实时报警
  - 可靠性与业务系统等同
- 统计
  - 反馈业务指标
  - 非实时（数小时~一天）
  - 灵活变化
- 90%以上数据项属于监控。需要专用监控系统

监控

- 监控系统
  - 极致的简单
  - **AttrAPI**：单一数值取样接口，易于添加，所见即所得
  - 数千监控项
- 统计系统
  - 极致的灵活
  - **OssLog**：日志汇总接口
  - 大日志量，数百统计项
  - **Hadoop**

- 在故障可被用户感知前排除它

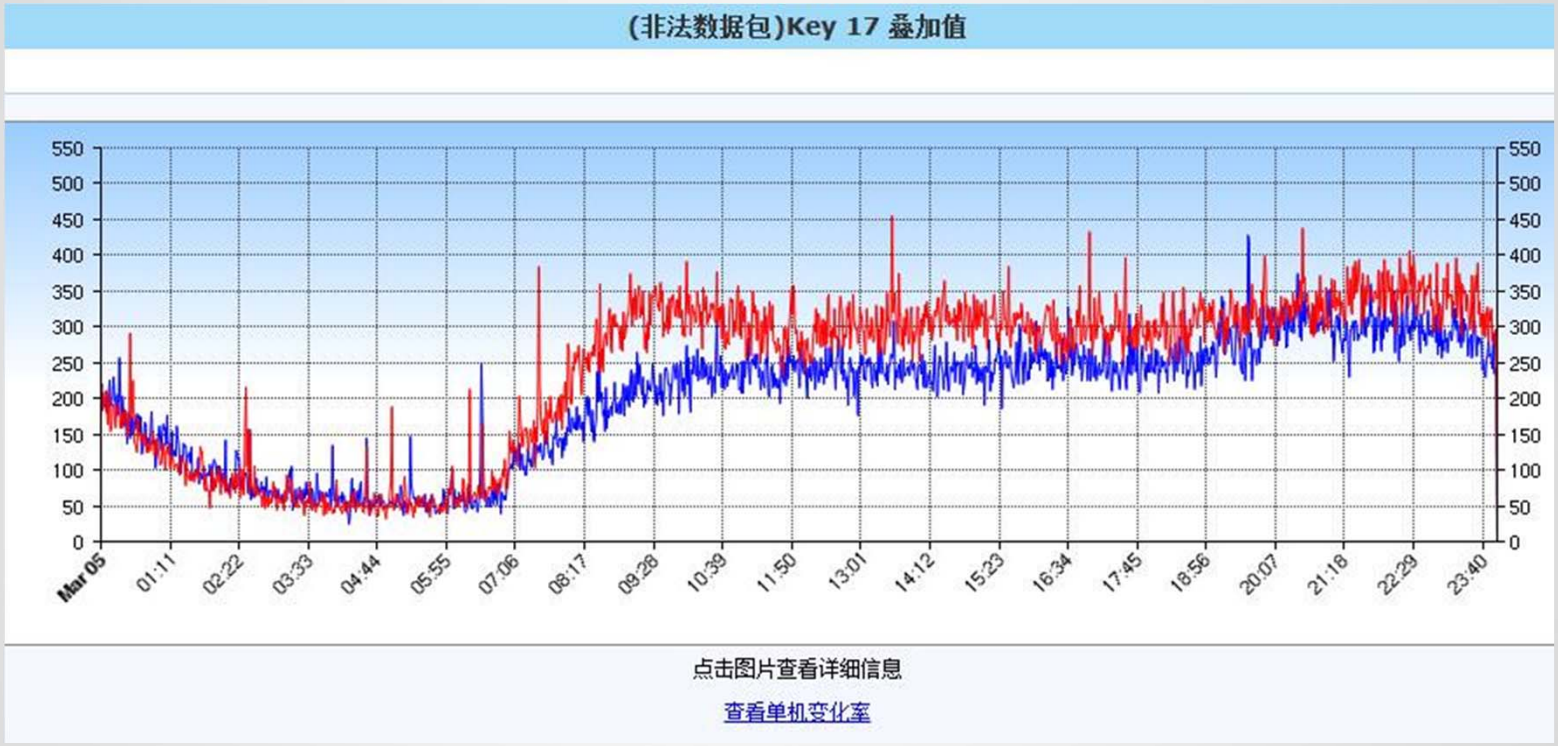
春晚时段监控曲线



LBS地图

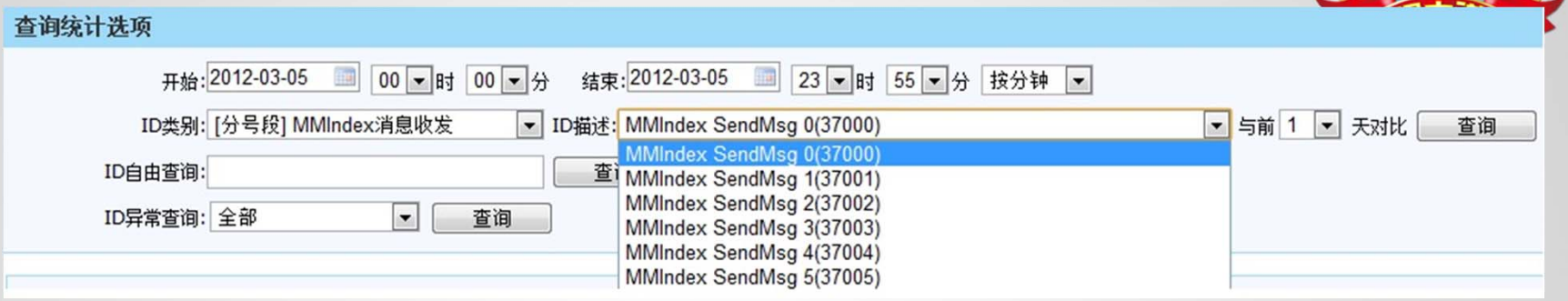
djt.qq.com

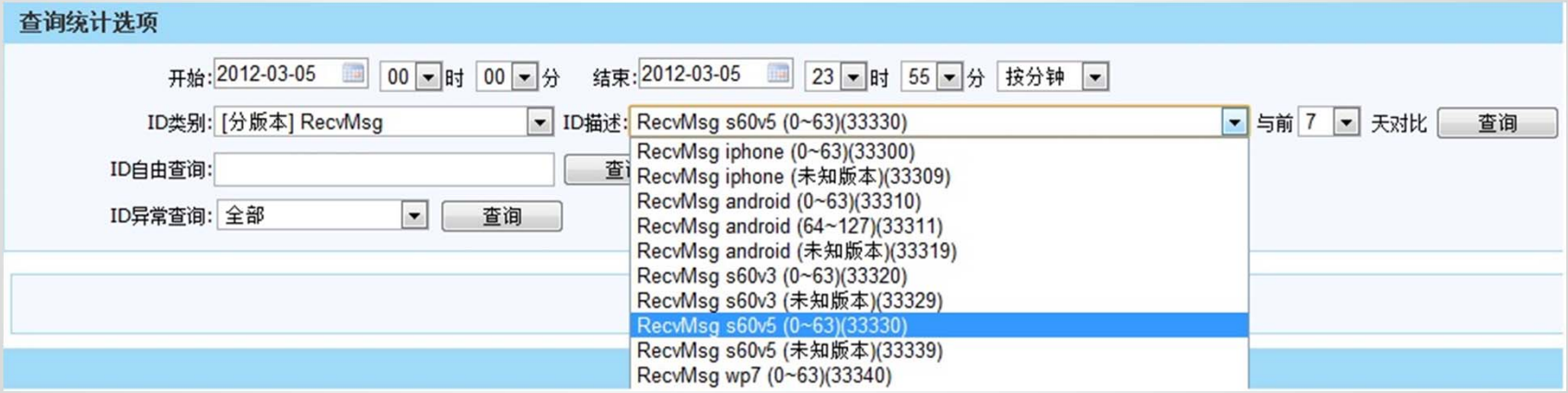
让监控更灵敏—捕捉异常



腾讯大讲堂 http://djt.qq.com

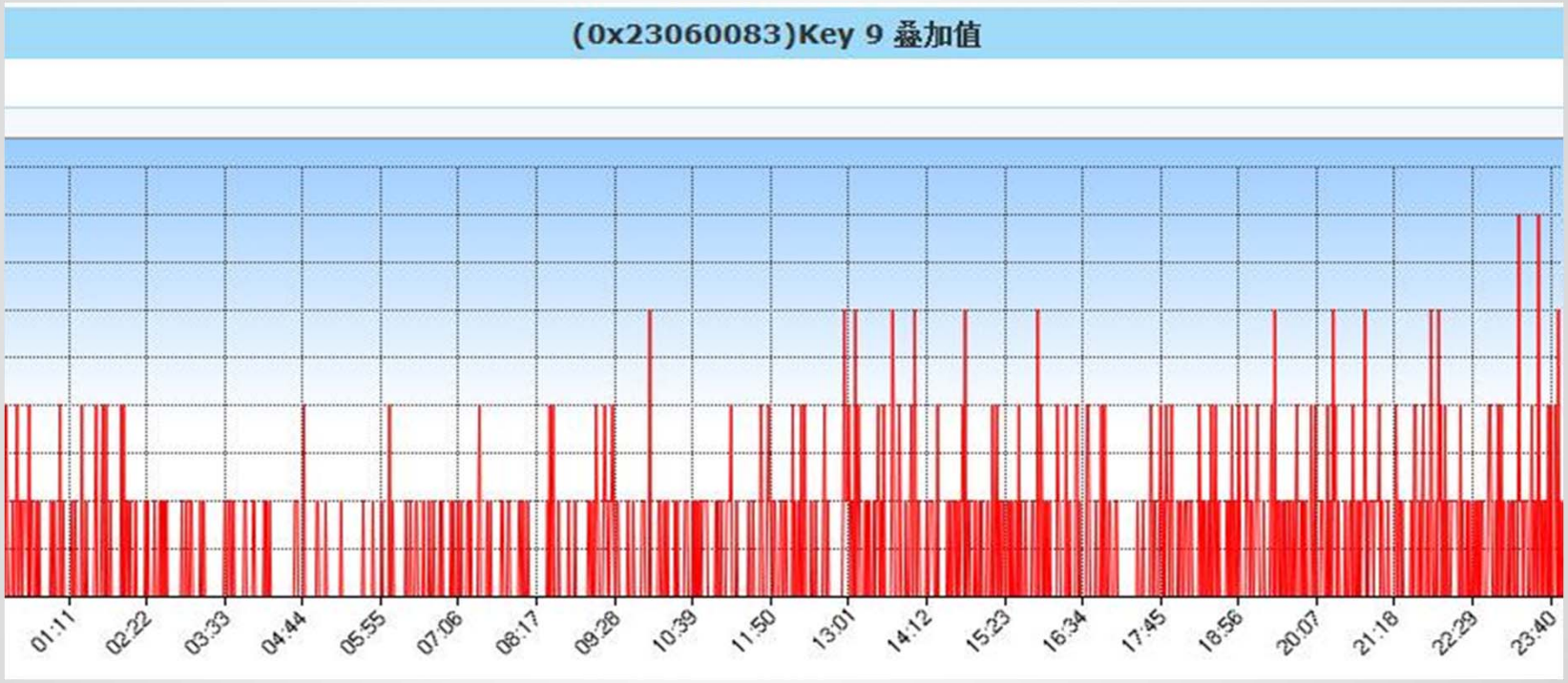
让监控更灵敏—分段监控


查询统计选项
开始:2012-03-05 00 时 00 分 结束:2012-03-05 23 时 55 分 按分钟
ID类别: [分号段] MMIndex消息收发
ID描述: MMIndex SendMsg 0(37000)
与前 1 天对比 查询
ID自由查询:
ID异常查询: 全部 查询
MMIndex SendMsg 0(37000)
MMIndex SendMsg 1(37001)
MMIndex SendMsg 2(37002)
MMIndex SendMsg 3(37003)
MMIndex SendMsg 4(37004)
MMIndex SendMsg 5(37005)


查询统计选项
开始:2012-03-05 00 时 00 分 结束:2012-03-05 23 时 55 分 按分钟
ID类别: [分版本] RecvMsg
ID描述: RecvMsg s60v5 (0~63)(33330)
与前 7 天对比 查询
ID自由查询:
ID异常查询: 全部 查询
RecvMsg iphone (0~63)(33300)
RecvMsg iphone (未知版本)(33309)
RecvMsg android (0~63)(33310)
RecvMsg android (64~127)(33311)
RecvMsg android (未知版本)(33319)
RecvMsg s60v3 (0~63)(33320)
RecvMsg s60v3 (未知版本)(33329)
RecvMsg s60v5 (0~63)(33330)
RecvMsg s60v5 (未知版本)(33339)
RecvMsg wp7 (0~63)(33340)

让监控更灵敏－灰度的利器

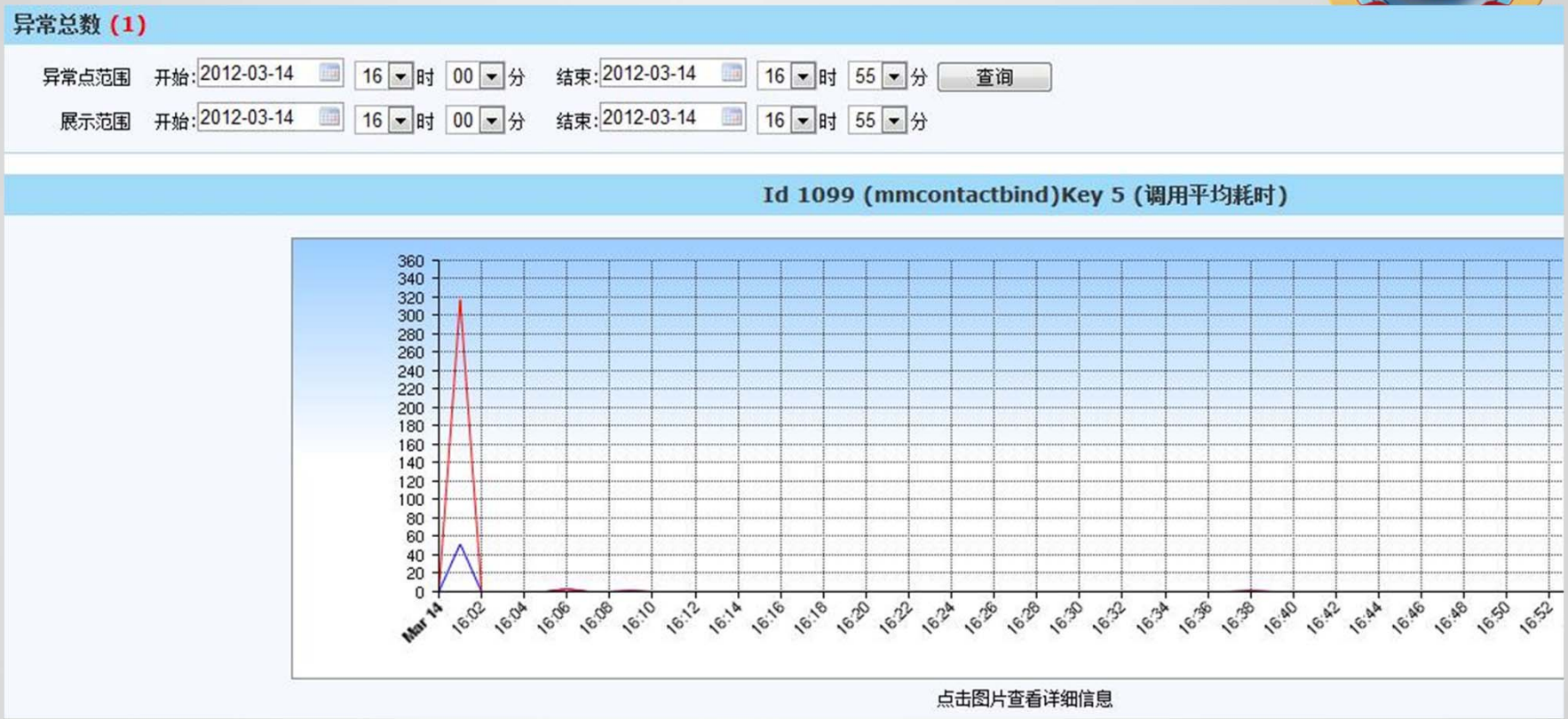
Example: Android Crash Report



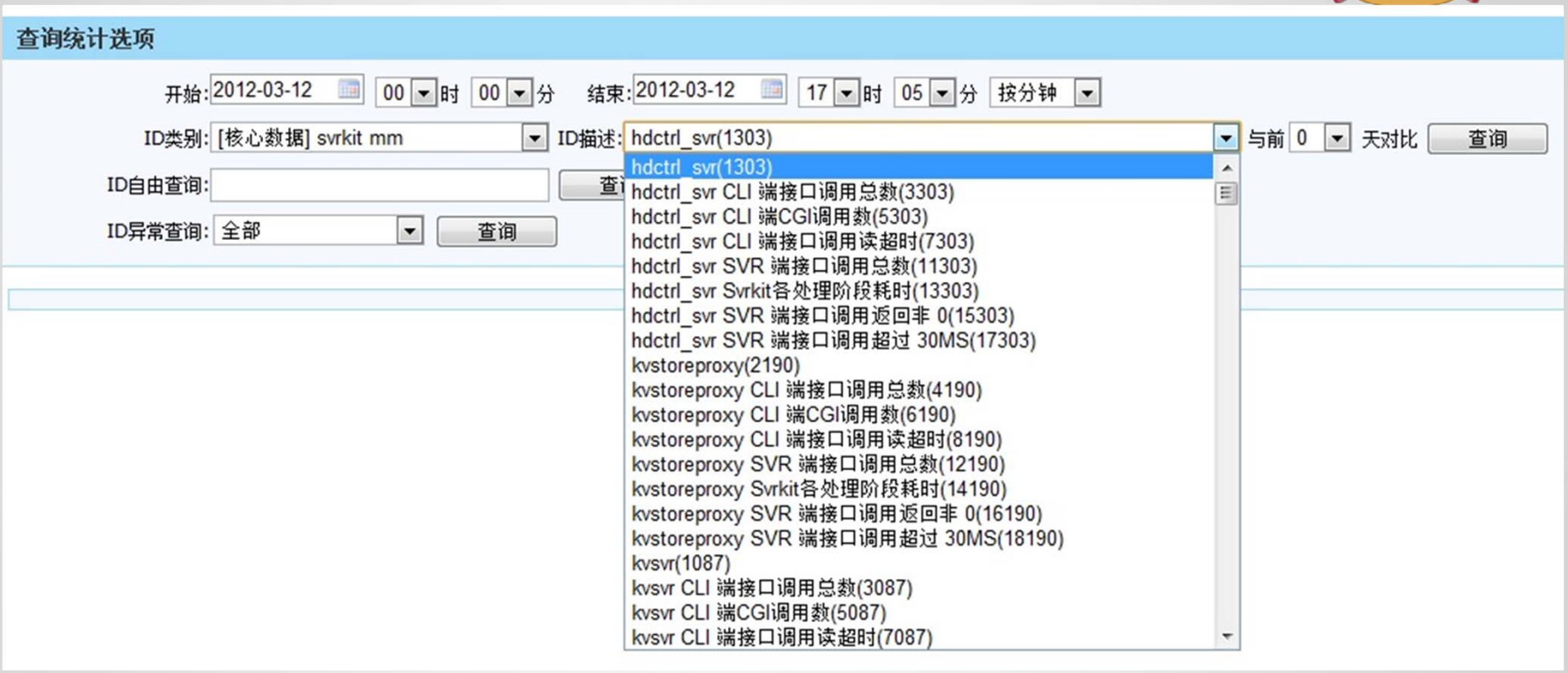
让监控更准确－监控点前移

	键值key	键值描述
☑	0	iphone平台收取消息延时0-1秒
☑	1	iphone平台收取消息延时1-2秒
☑	2	iphone平台收取消息延时2-3秒
☑	3	iphone平台收取消息延时3-5秒
☑	4	iphone平台收取消息延时5-10秒
☑	5	iphone平台收取消息延时10-20秒
☑	6	iphone平台收取消息延时20-30秒
☑	7	iphone平台收取消息延时30-60秒
☑	8	iphone平台收取消息延时60-120秒
☑	9	iphone平台收取消息延时120-180秒
☑	10	iphone平台收取消息延时180-300秒
☑	11	iphone平台收取消息延时300-600秒
☑	12	iphone平台收取消息延时600-1200秒
☑	13	iphone平台收取消息延时1200-1800秒
☑	14	iphone平台收取消息延时1800-3600秒
☑	15	iphone平台收取消息延时大于3600秒

自动报警


异常总数 (1)
异常点范围 开始:2012-03-14 16 时 00 分 结束:2012-03-14 16 时 55 分 查询
展示范围 开始:2012-03-14 16 时 00 分 结束:2012-03-14 16 时 55 分
Id 1099 (mmcontactbind)Key 5 (调用平均耗时)
360 340 320 300 280 260 240 220 200 180 160 140 120 100 80 60 40 20 0
Mar 14 16:02 16:04 16:06 16:08 16:10 16:12 16:14 16:16 16:18 16:20 16:22 16:24 16:26 16:28 16:30 16:32 16:34 16:36 16:38 16:40 16:42 16:44 16:46 16:48 16:50 16:52
点击图片查看详细信息

把监控嵌入基础框架


查询统计选项
开始:2012-03-12 00 时 00 分
结束:2012-03-12 17 时 05 分 按分钟
ID类别: [核心数据] svrkit mm
ID描述: hdctrl_svr(1303)
与前 0 天对比
查询
ID自由查询:
ID异常查询: 全部
查询
hdctrl_svr(1303)
hdctrl_svr CLI 端接口调用总数(3303)
hdctrl_svr CLI 端CGI调用数(5303)
hdctrl_svr CLI 端接口调用读超时(7303)
hdctrl_svr SVR 端接口调用总数(11303)
hdctrl_svr Svrkit各处理阶段耗时(13303)
hdctrl_svr SVR 端接口调用返回非 0(15303)
hdctrl_svr SVR 端接口调用超过 30MS(17303)
kvstoreproxy(2190)
kvstoreproxy CLI 端接口调用总数(4190)
kvstoreproxy CLI 端CGI调用数(6190)
kvstoreproxy CLI 端接口调用读超时(8190)
kvstoreproxy SVR 端接口调用总数(12190)
kvstoreproxy Svrkit各处理阶段耗时(14190)
kvstoreproxy SVR 端接口调用返回非 0(16190)
kvstoreproxy SVR 端接口调用超过 30MS(18190)
kvsvr(1087)
kvsvr CLI 端接口调用总数(3087)
kvsvr CLI 端CGI调用数(5087)
kvsvr CLI 端接口调用读超时(7087)

总结

- 三位一体
  - 产品的精准
  - 项目的敏捷
  - 技术的支撑
- 剥离复杂，让剩下的更简单
  - 协议
  - 容灾
  - 轻重
  - 监控
- 一些原则
  - 大系统小做
  - 面向最坏的思考，柔性可用
  - 分而治之

最后，让剩下的更简单

- 摇一摇 & 漂流瓶，一周完成
- 3个月30个内部发布
- 每天20个后台变更
- 99.95%的可用性

未来的技术挑战

- **99.99%**
- 面向**10**倍的架构提升
- 完全的**IDC**容灾

技术的追求

- 其疾如风，其徐如林，侵掠如火，不动如山

Q & A